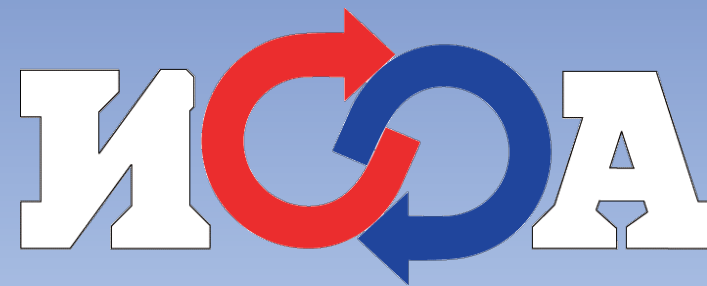


Применение машинного обучения для создания адаптивных алгоритмов распознавания угроз конвективных опасных явлений

Шершакова Анна Олеговна
МГТУ им. Н.Э. Баумана,
ИФА им. А.М. Обухова РАН

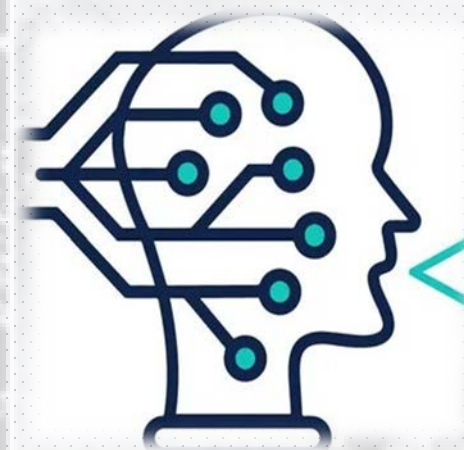


Конференция «Турбулентность,
динамика атмосферы и климата»

ВВЕДЕНИЕ

Машинное обучение – одно из направлений искусственного интеллекта, нацеленное на создание моделей, которые могут обучаться и развиваться на основе получаемых эмпирических данных данных.

Использование методов машинного обучения может способствовать выявлению скрытых особенностей и закономерностей протекания процесса, учет которых в перспективе позволит повысить качество прогнозирования.



ОБЪЕКТ ИССЛЕДОВАНИЯ – ВОДЯНЫЕ СМЕРЧИ

За последние несколько лет участились случаи возникновения черноморских смерчей в опасной близости к побережью. Для минимизации возможного ущерба от выхода смерчей на сушу важно своевременно распознать угрозу их образования.



22.08.2017, Сочи



04.07.2021, Сочи



11.07.2023, Туапсе

ЦЕЛИ И ЗАДАЧИ

ЦЕЛЬ

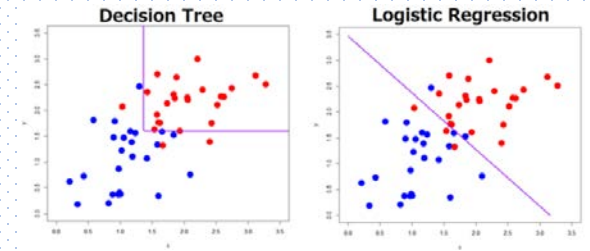
Построение адаптивных алгоритмов распознавания угроз возникновения опасных явлений

1. Исследование возможности применения методов интеллектуального анализа данных и машинного обучения для наукастинга

2. Разработка эффективной методики обработки и анализа динамических характеристик облачных ячейках в Черноморском регионе.

3. Построение ряда математических моделей классификации ячеек по степени опасности формирования из них водяных смерчей и оценка эффективности моделей.

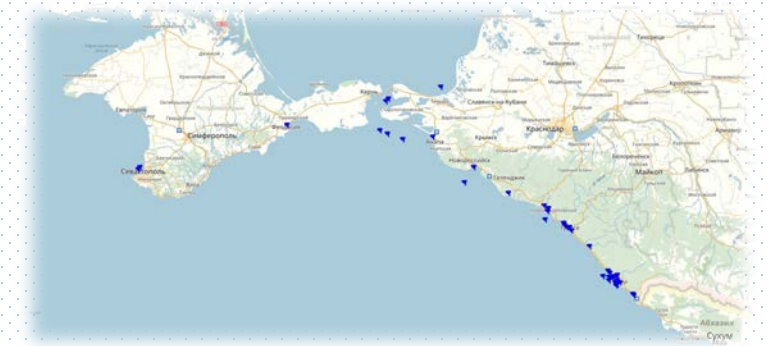
4. Проверка качества прогнозных свойств моделей на новых данных в «квазиоперативном режиме»



Пример работы моделей машинного обучения по разделению выборки данных на два класса в двумерном признаковом пространстве

$$\begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_l) & \dots & f_n(x_l) \end{pmatrix} \xrightarrow{y} \begin{pmatrix} y_1 \\ \dots \\ y_l \end{pmatrix} \xrightarrow{\mu} a$$
$$\begin{pmatrix} f_1(x'_1) & \dots & f_n(x'_1) \\ \dots & \dots & \dots \\ f_1(x'_k) & \dots & f_n(x'_k) \end{pmatrix} \xrightarrow{a} \begin{pmatrix} a(x'_1) \\ \dots \\ a(x'_k) \end{pmatrix}$$

Задача классификации. Машинное обучение



Положение смерчей в акватории Черного моря в 2016 г.

Процесс Data Science

1. Назначение цели исследования

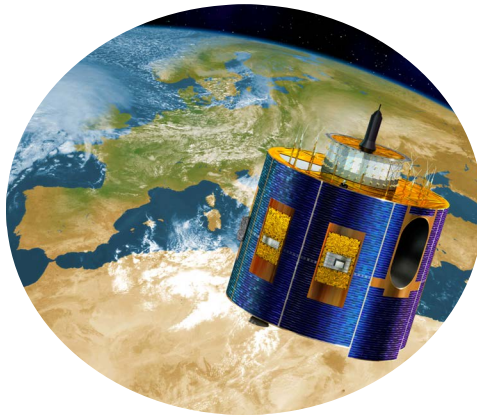
- построение моделей распознавания облаков с водяными смерчами с применением интеллектуального анализа данных и машинного обучения для последующего наукастинга этого природного явления



2. Сбор данных

Спутник Meteosat Second Generation – 10 (MSG3)

- набор характеристик облачных ячеек;
- перечень ячеек, из которых возникали смерчи.

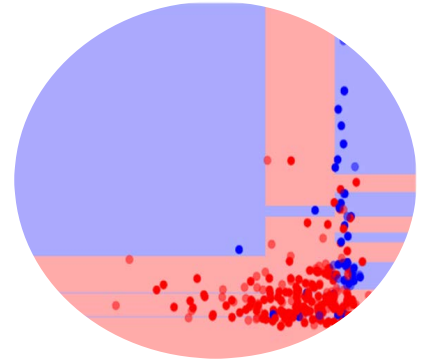


3. Подготовка и исследование данных

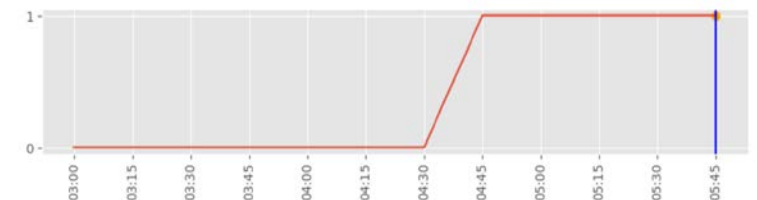
- первичная подготовка данных
- построение предикторов прогностической модели
- анализ корреляционных зависимостей в данных

4. Моделирование Машинное обучение - задача классификации

- линейная классификация
- логистическая регрессия
- модель опорных векторов
- стохастический градиентный спуск
- метод ближайшего соседа
- метод решающих деревьев
- случайный лес
- оптимизированный градиентный бустинг
- адаптивный бустинг



5. Анализ результатов, отчет



Исходные данные

- набор характеристик облачных ячеек (349147 измерений)
- перечень ячеек, из которых возникали смерчи (87)

Первичная подготовка

- анализ
- удаление неинформативных параметров
- выявление и обработка выбросов

data
все ячейки

	num	dt	fnum	fdt	bdt	severity	confidence_level	categ	categ_age	ctype	...	btd5q1	btd5q2	btd5ratio	lon_l
id_all															
1	1	201405310000	-1	-1	201405310000	0	0	5	0	0.0	...	-1.455109	-1.100891	0.0	
2	2	201405310000	-1	-1	201405310000	0	0	5	0	0.0	...	-1.478912	-1.397186	0.0	
3	3	201405310000	-1	-1	201405310000	0	0	5	0	0.0	...	-1.572784	-1.025024	0.0	
4	4	201405310000	-1	-1	201405310000	0	0	5	0	0.0	...	-0.859680	-0.531189	0.0	
5	5	201405310000	-1	-1	201405310000	0	0	5	0	0.0	...	-2.541321	-2.458435	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
349143	133	201612030900	72	201612030845	201612030715	0	0	2	0	0.0	...	0.009293	0.131775	0.0	
349144	134	201612030900	-1	-1	201612030900	0	0	3	0	0.0	...	-1.004745	-0.599472	0.0	
349145	135	201612030900	-1	-1	201612030900	0	0	3	0	0.0	...	-1.173050	-0.864151	0.0	
349146	136	201612030900	73	201612030845	201612030815	0	0	3	45	0.0	...	-0.138947	0.007858	0.0	
349147	137	201612030900	75	201612030845	201612030845	0	0	3	15	0.0	...	-0.292984	-0.190933	0.0	

349147 rows x 79 columns

Часть DataFrame со всеми параметрами облачных ячеек(data)

waterspoutCells
смерчевые ячейки

	num	dt	bnum	bdt
id_water				
1	74	201405310630	152.0	2.014053e+11
2	19	201406100945	17.0	2.014061e+11
3	58	201406131330	59.0	2.014061e+11
4	68	201406140930	47.0	2.014061e+11
5	56	201406141000	47.0	2.014061e+11
...	...	...	...	...
83	32	201609180430	22.0	2.016092e+11
84	39	201609180845	12.0	2.016092e+11
85	33	201609181030	37.0	2.016092e+11
86	46	201609191945	71.0	2.016092e+11
87	94	201612030530	86.0	2.016120e+11

87 rows x 4 columns

DataFrame со всеми смерчевыми
облачными
ячейками(waterspoutCells)

Полная база данных облачных ячеек (DataFrame)

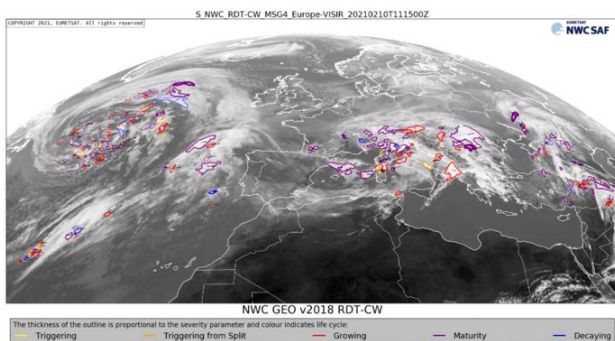
На первом этапе работы с данными из многочисленных файлов табличного формата (csv), в которых находились значения параметров облачных ячеек, была сфорсирована одну информативную таблицу **data** формата Pandas DataFrame.

Статистика:

- 349147 строк - измерения облачных ячейки
- 79 столбцов - характеристики облачных ячеек



Анализ динамики характеристи к ячеек. Meteo-France



Пример представления результатов работы модуля RDT- Rapid Developing

Thunderstorm

Подход к классификации облачных ячеек Meteo-France

Входные параметры:

- минимальное значение параметра t_{min} на протяжении времени жизни ячейки (далее просто ВЖ);
- первый максимум параметра col_rate на ВЖ;
- второй максимум параметра col_rate на ВЖ;
- третий максимум параметра col_rate на ВЖ;
- четвертый максимум параметра col_rate на ВЖ;
- пятый максимум параметра col_rate на ВЖ;
- минимальное значение параметра col_rate на ВЖ;
- является ли значение параметра col_rate положительным на ВЖ (да/нет);
- является ли значение параметра col_rate отрицательным на ВЖ (да/нет);
- является ли максимальное значение параметра col_rate положительным, а минимальное – отрицательным на ВЖ (да/нет);
- максимальная скорость изменения параметра $tavg$ (под скоростью понимается разность $tavg$ родителя ячейки на предыдущем цикле сканирования и $tavg$ ячейки, деленная на частоту сканирования (1/4 часа));
- максимальная скорость изменения параметра $tavg_st$;
- максимальное значение разности параметров $tavg$ и $tmin$ на ВЖ;
- минимальное значение разности параметров $tavg$ и $tmin$ на ВЖ.

✓
0
сек.

[64] history_2[0]

	num	dt	fnum	fdt	bdt	surface	gaxe_km	paxe_km	tth	tmin	...	indikator
id_all												
5455	79	201405310615	152	201405310600	201405310600	378.125458	29.532776	19.663921	4.0	-9.481390	...	1
5675	74	201405310630	79	201405310615	201405310600	681.123657	54.059612	23.746243	-4.0	-38.595558	...	2

2 rows × 60 columns

История смерчевой ячейки глубины 2

✓
0
сек.

[66] history_3[0]

	num	dt	fnum	fdt	bdt	surface	gaxe_km	paxe_km	tth	tmin	...	indikator
id_all												
5312	152	201405310600	-1	-1	201405310600	294.976440	24.659275	17.268166	5.0	-5.368598	...	1
5455	79	201405310615	152	201405310600	201405310600	378.125458	29.532776	19.663921	4.0	-9.481390	...	1
5675	74	201405310630	79	201405310615	201405310600	681.123657	54.059612	23.746243	-4.0	-38.595558	...	2

3 rows × 60 columns

История смерчевой ячейки глубины 3

✓
0
сек.

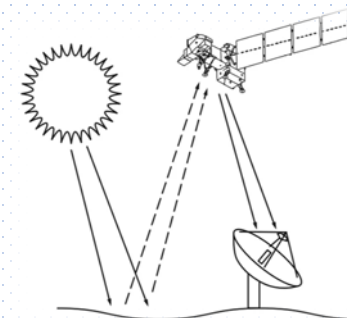
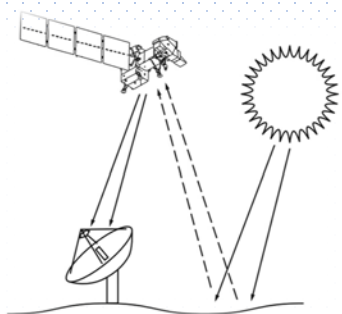
▶ history_4[0]

	num	dt	fnum	fdt	bdt	surface	gaxe_km	paxe_km	tth	tmin	...	indikator
id_all												
19622	56	201406131200	107	201406131145	201406131130	1274.734375	55.050373	33.396027	-8.0	-39.652260	...	1
19778	43	201406131215	56	201406131200	201406131130	1758.928345	68.872681	36.693081	-6.0	-42.935463	...	1
19968	84	201406131230	43	201406131215	201406131130	2312.053467	83.298141	39.204517	-6.0	-47.310066	...	1
20121	91	201406131245	84	201406131230	201406131130	2452.562988	88.132080	41.157497	-7.0	-46.408562	...	1
20282	96	201406131300	91	201406131245	201406131130	4305.884766	113.971085	62.354950	-5.0	-45.227943	...	1
20420	89	201406131315	96	201406131300	201406131130	4919.393066	118.545242	75.680634	-5.0	-42.654991	...	1
20523	58	201406131330	89	201406131315	201406131130	7181.908691	148.935745	134.089096	-5.0	-40.187416	...	2

7 rows × 60 columns

История смерчевой ячейки глубины 4 и более

История смерчевой ячейки -
структура, хранящую все
измерения облачной ячейки от
момента ее формирования до
реализции в смерч



История условно безопасной ячейки - структура, хранящую все измерения облачной

history_all_2[10]

ячейки от момента ее формирования до исчезновения

	num	dt	fnum	fdt	bd	surface	gaxe_km	paxe_km	tth	tmin	...	tavg_st	tmin_st	tth_st	grad_peri_st	surface_st	indikator
id_all																	
309102	37	201607191115	38	201607191100	201607191100	449.316376	35.387482	15.188237	-2.0	-14.362249	...	-11.747620	-14.362249	-8.0	0.945914	173.028412	0
309171	35	201607191130	37	201607191115	201607191100	904.632202	56.358669	28.912245	-2.0	-10.897741	...	-7.535774	-10.897741	-4.0	0.960669	277.703094	0

2 rows × 60 columns

История безопасной ячейки глубины 2

history_all_3[24]

	num	dt	fnum	fdt	bd	surface	gaxe_km	paxe_km	tth	tmin	...	tavg_st	tmin_st	tth_st	grad_peri_st	surface_st	indikator
id_all																	
130329	70	201504200330	79	201504200315	201504200230	3919.307129	142.599686	68.898201	-33.0	-40.727242	...	-36.369450	-40.727242	-34.0	0.538490	2645.336670	0
130425	77	201504200345	70	201504200330	201504200230	2824.068115	86.413574	58.055054	-32.0	-38.073784	...	-34.322277	-38.073784	-32.0	0.429901	2824.068115	0
130517	78	201504200400	77	201504200345	201504200230	1806.234619	79.299339	43.704208	-32.0	-38.073784	...	-34.265793	-38.073784	-32.0	0.431585	1806.234619	0

3 rows × 60 columns

История безопасной ячейки глубины 3

history_all_4[136]

	num	dt	fnum	fdt	bd	surface	gaxe_km	paxe_km	tth	tmin	...	tavg_st	tmin_st	tth_st	grad_peri_st	surface_st	indikator
id_all																	
282589	160	201605241345	198	201605241330	201605241115	452.705933	25.428726	23.526899	-29.0	-40.187416	...	-36.785763	-40.187416	-34.0	1.027718	254.774521	0
282869	198	201605241400	160	201605241345	201605241115	3278.868164	116.422684	55.580593	-19.0	-42.654991	...	-39.187447	-42.654991	-36.0	1.089233	494.718201	0
283106	208	201605241415	198	201605241400	201605241115	3843.203125	120.591316	55.356842	-19.0	-44.070641	...	-40.852386	-44.070641	-38.0	1.323517	321.710571	0
283265	132	201605241430	208	201605241415	201605241115	3893.098633	128.583511	52.692745	-19.0	-42.375832	...	-39.315525	-42.375832	-36.0	0.968413	469.455963	0
283468	106	201605241445	132	201605241430	201605241115	3760.054932	128.189651	50.772533	-19.0	-40.187416	...	-36.657570	-40.187416	-34.0	1.056864	501.063629	0
283729	147	201605241500	106	201605241445	201605241115	1673.580444	109.036682	41.244545	-25.0	-38.858070	...	-35.058926	-38.858070	-32.0	1.030791	295.936096	0
283941	126	201605241515	147	201605241500	201605241115	475.344299	36.970181	16.172768	-25.0	-35.527184	...	-32.712585	-35.527184	-29.0	1.045633	267.455170	0

7 rows × 60 columns

История безопасной ячейки глубины 4 и более

Предикторы. Динамические характеристик и ячеек

Правила построения предикторов прогностической модели

(максимальная рассматриваемая глубина истории – 90 минут (7 измерений)):

общие (для всех параметров)

- максимальная скорость изменения параметра (1, 2, 3 максимумы)
- максимальное значение параметра (1, 2 максимумы)
- минимальное значение параметра

специальные

- максимальное значение разности параметров `surface` И `surface_st`
- максимальное значение отношения `pace_km/gaxe_km`
- максимальное значение разности параметров `tthst` И `tavgst`
- максимальное значение разности параметров `tavg` И `tmin`
- минимальное значение разности параметров `tavg` И `tmin`
- максимальное значение разности параметров `tavgst` И `tminst`
- максимальное значение разности параметров `tth` И `tavg`
- является ли значение параметра `col_rate` положительным (да (1)/нет (0))
- является ли значение параметра `col_rate` отрицательным (да (1)/нет (0))
- является ли максимальное значение параметра `col_rate` положительным,
а минимальное – отрицательным (да (1)/нет (0))

«Динамическая» обучающая выборка

	id_all_last	num_last	dt		bdt	len_history	indikator_danger	max_razn_surface.surface_st	max_razn_tth_st.tavg_st	max_razn_tavg.tmin
0	116399.0	211.0	2.015040e+11	2.015040e+11		3.0	0.0	676.407898	2.873110	9.031256
1	180336.0	119.0	2.015063e+11	2.015063e+11		3.0	0.0	347.842285	5.009892	10.266503
2	251917.0	108.0	2.015082e+11	2.015082e+11		3.0	0.0	70.408447	4.191566	4.051037
3	207452.0	43.0	2.015063e+11	2.015063e+11		3.0	0.0	385.216324	4.615965	12.773790
4	279906.0	151.0	2.016051e+11	2.016051e+11		3.0	0.0	1533.224640	6.519543	14.621139
...	...	...	...	...	...	...	...	...	...	...
269	235404.0	40.0	2.015073e+11	2.015073e+11		3.0	0.0	402.146789	2.958834	5.478141
270	134132.0	100.0	2.015042e+11	2.015042e+11		3.0	0.0	640.460175	3.846264	9.041886
271	147859.0	210.0	2.015062e+11	2.015062e+11		3.0	0.0	361.240570	4.517481	8.518179
272	189876.0	125.0	2.015063e+11	2.015063e+11		3.0	1.0	701.091232	3.224941	11.815212
273	167338.0	17.0	2.015063e+11	2.015063e+11		3.0	0.0	338.571625	2.948849	7.412300

274 rows × 387 columns

grad_peri_st_max_speed_2	surface_st_max_1	surface_st_min_1	surface_st_max_2	surface_st_min_2	surface_st_max_speed_1	surface_st_max_3	surface_st_max_speed_2
-0.227086	331.393738	279.682983	327.557465	327.557465	0.171174	279.682983	0.011712
-0.113545	231.453918	87.073563	87.074341	87.074341	1.658119	87.073563	0.000009
-0.145472	327.556519	117.259460	257.774384	257.774384	-0.213038	117.259460	-0.545108
-0.005592	115.249100	76.874313	114.935127	114.935127	0.499189	76.874313	-0.331150
0.011562	158.301117	78.343727	157.718903	157.718903	1.013166	78.343727	0.003691
...	...	...	...	...	...	...	...
-0.193632	644.313721	483.163177	643.494202	643.494202	0.001274	483.163177	-0.250112
NaN	351.042084	127.678482	255.129913	255.129913	-0.273221	127.678482	-0.499555
0.255571	197.916290	98.805664	131.725510	131.725510	0.502490	98.805664	0.333178
-0.419989	575.977722	255.745316	256.288452	256.288452	1.247381	255.745316	0.002124
-0.058567	270.827881	180.555450	225.600586	225.600586	0.249481	180.555450	-0.333320

Пример части DataFrame со всеми предикторами облачных ячеек с глубиной истории 3

Масштабность моделирования

Разработка оптимального алгоритма
подготовки данных

225 моделей

9 методов машинного обучения

5 типов формирования обучающей выборки
(по глубине истории и количеству
предикторов)

5 способов обработки корреляционных
зависимостей для минимизации признакового
пространства

```
name = "test_" + history_type + "for_all_find_" + clf + "_" + corr_type
```

где *name* — итоговое имя модели;

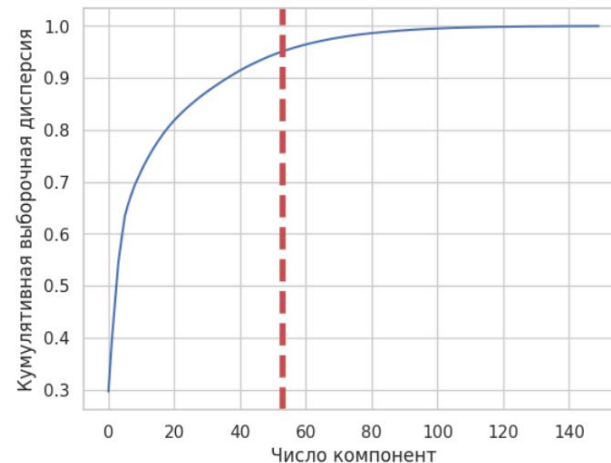
history_type = {2,3,3_small,4,4_small} — тип глубины истории и количества
предикторов;

clf = {lin,log,metric,tree,forest,XGB,SVC,ADA,SGD} — метод машинного
обучения;

corr_type = {classic,del_right,del_back,pca,pca_del} — методика обработки
корреляционных зависимостей

Методы машинного обучения

- линейная классификация (lin)
- логистическая регрессия (log)
- модель опорных векторов (svs)
- стохастический градиентный спуск (SGD)
- метод ближайшего соседа (metric)
- метод решающих деревьев (tree)
- случайный лес (forest)
- оптимизированный градиентный бустинг (XGB)
- адаптивный бустинг (ADA)



Пример графика зависимости доли выборочной дисперсии
от числа главных компонент

Тип обучающей выборки

- предикторы на глубине историй 2
- предикторы на глубине историй 3
- предикторы на глубине историй 3
(сокращенный набор)
- предикторы на глубине историй 4 и более
- предикторы на глубине историй 4 и более
(сокращенная версия)

Обработка зависимостей

- без очистки от корреляционных
зависимостей (classic)
- удаление зависимых параметров
(2)(del_right, del_back)
- метод главных компонент (pca)
- удаление параметров + метод главных
компонент (pca_del)

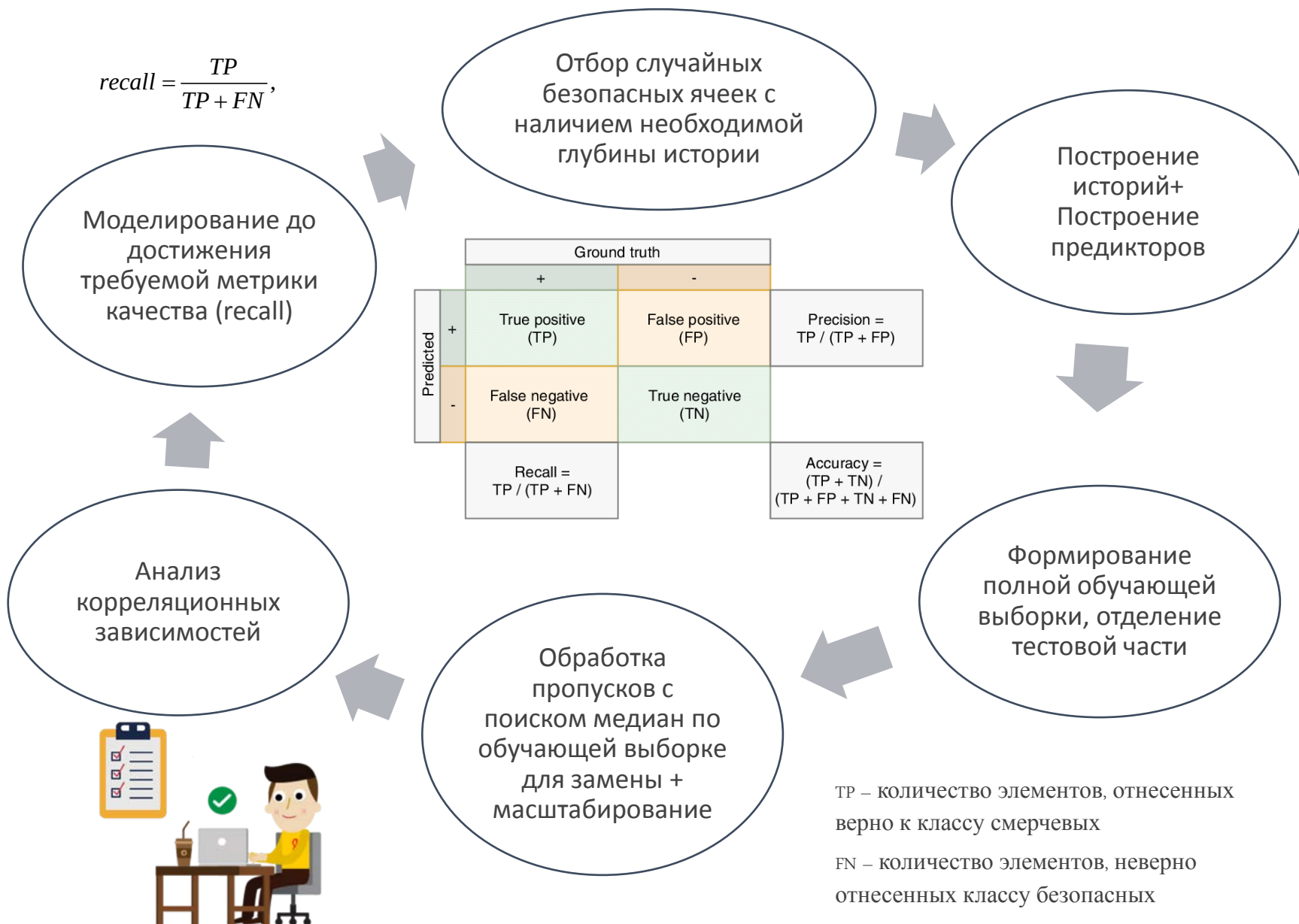
Цикличность поиска оптимальной модели

На результат оказывает влияние очень много факторов, поэтому обучение и моделирование производилось в цикле до достижения метрики качества алгоритма

В случае достижения метрики качества сохраняем в базу данных:

- обучающую выборку
- параметры масштабирования и пары: предиктор с пропусками и медиана
- параметры PCA модели или список удаляемых параметров (в зависимости от методики)
- параметры модели

$$recall = \frac{TP}{TP + FN},$$

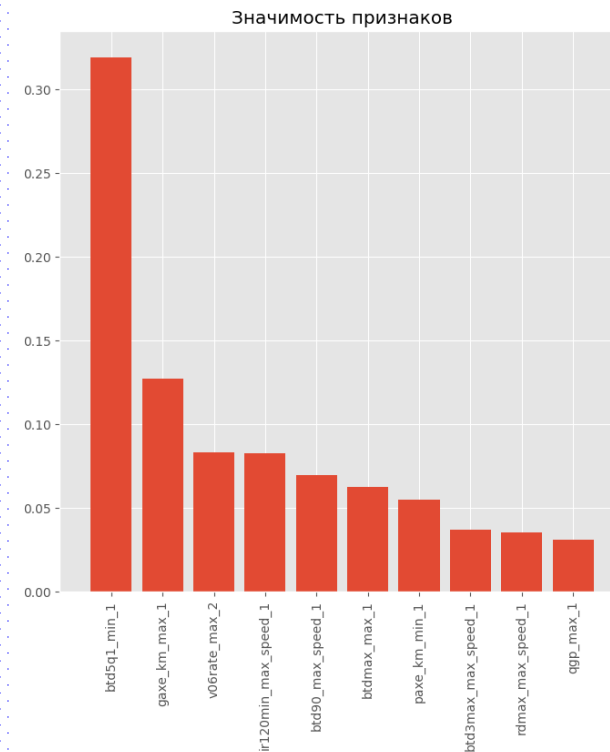


Первичный результат моделирования

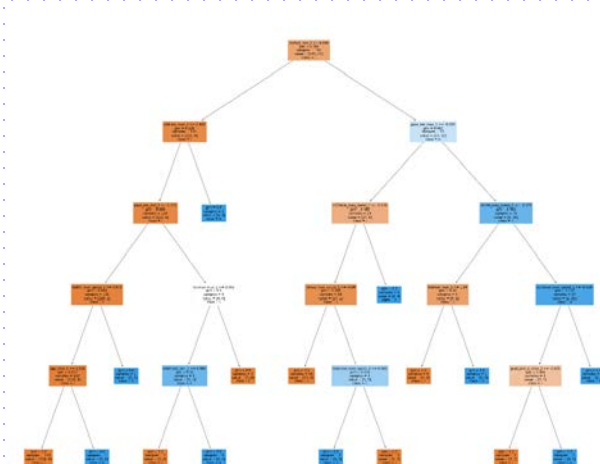
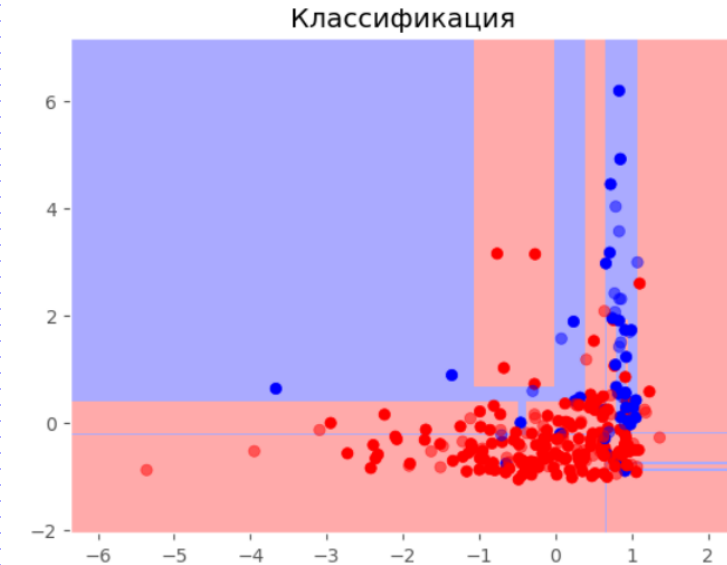
Получено 225 моделей классификации облачных ячеек

Требуется анализ!

Оценить характеристики для каждого типа машинного обучения, выявляли подходящий тип обработки данных и т.д.



Первые 3 значимых признаков ['btd5q1_min_1', 'gaxe_km_max_1', 'v06rate_max_2'] из 257
Классификация по 2 параметрам: Index(['btd5q1_min_1', 'gaxe_km_max_1'], dtype='object')
Recall = 0.55
Classification score was: 0.7875



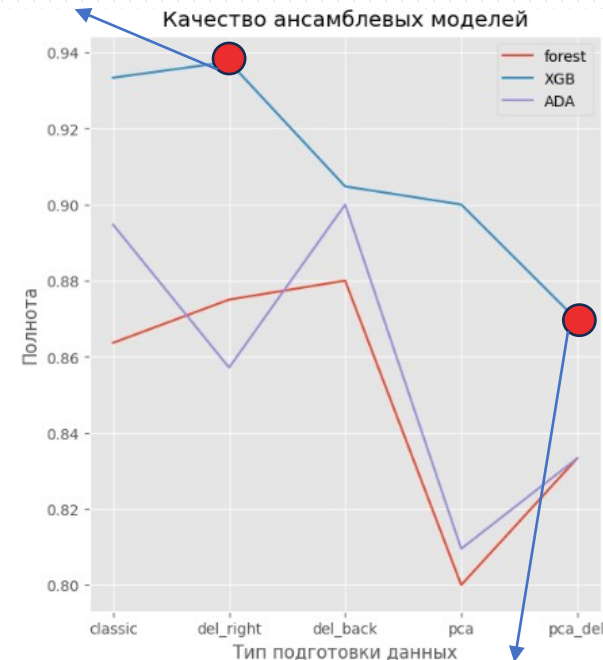
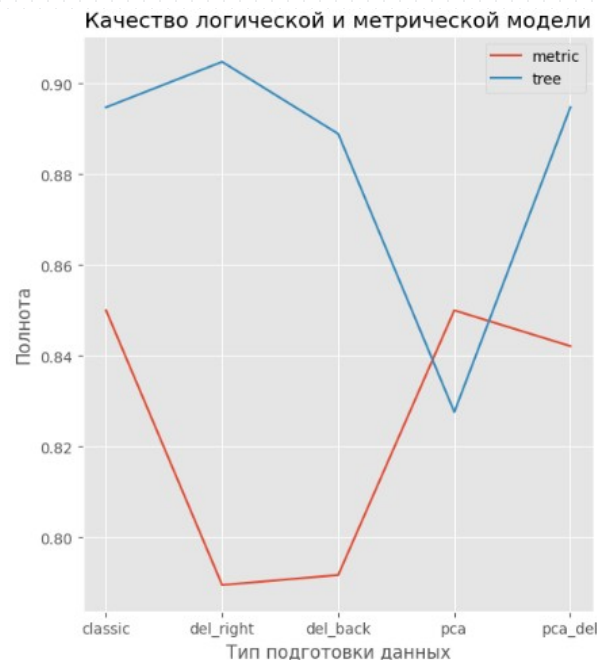
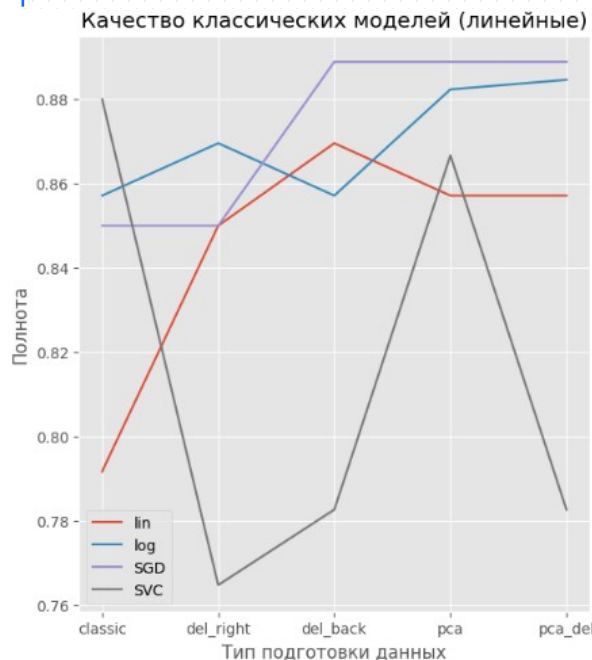
АНАЛИЗ

Зависимость качества классификации от типа подготовки данных для групп моделей. Истории глубиной 2

Типы подготовки данных. Сокращения.

- classic - без очистки от корреляционных зависимостей
- del_right, del_back - удаление зависимых параметров
- pca - метод главных компонент
- pca_del - удаление параметров + метод главных компонент

Модель оптимизированного градиентного бустинга (XGB) с удалением коррелирующих предикторов: полнота 94%



`name = "test_" + history_type + "for_all_find_" + clf + "_" + corr_type`

где `name` — итоговое имя модели;

`history_type = {2, 3, 3_small, 4, 4_small}` — тип глубины историй и количества предикторов;

`clf = {lin, log, metric, tree, forest, XGB, SVC, ADA, SGD}` — метод машинного обучения;

`corr_type = {classic, del_right, del_back, pca, pca_del}` — методика обработки корреляционных зависимостей

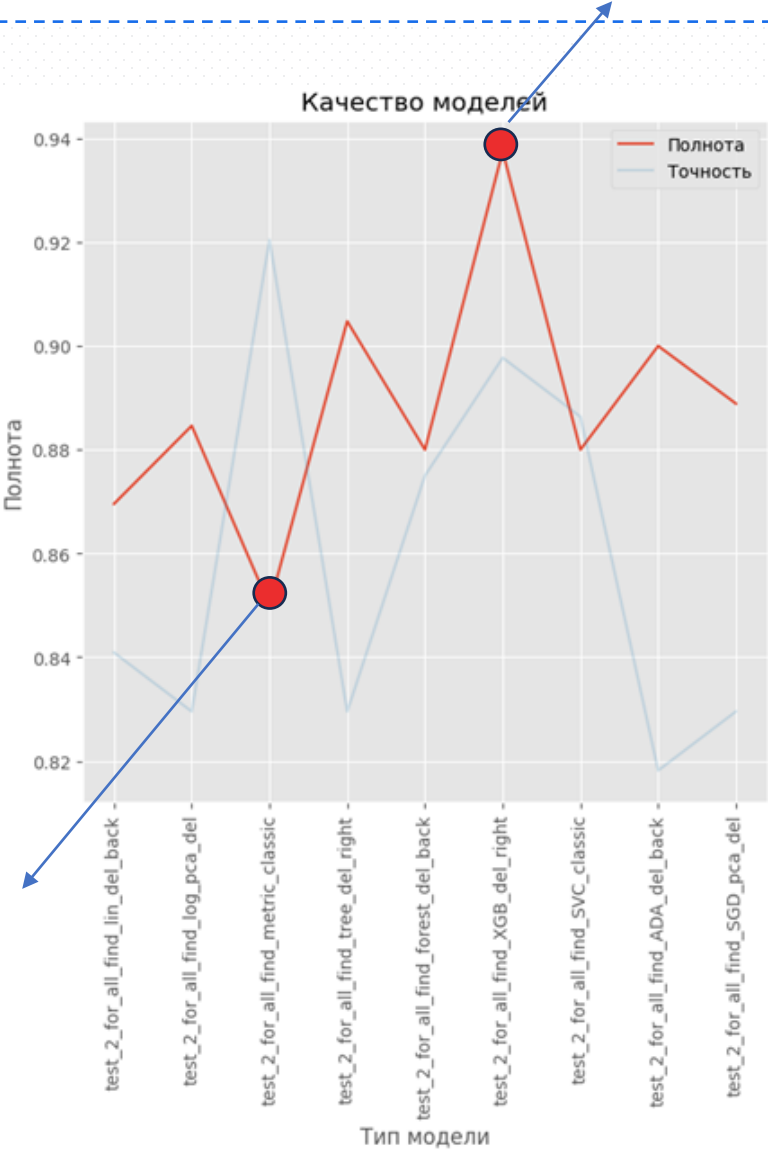
Модель оптимизированного градиентного бустинга (XGB) с удалением коррелирующих предикторов и применением метода главных компонент: полнота 87%

Модель оптимизированного градиентного бустинга
(XGB) с удалением коррелирующих предикторов:
полнота 94%, точность 90%

РЕЗУЛЬТАТЫ. ИСТОРИИ ГЛУБИНЫ 2

	pod	accur
test_2_for_all_find_XGB_del_right	0.93750	0.897727
test_2_for_all_find_tree_del_right	0.90476	0.829545
test_2_for_all_find_ADA_del_back	0.90000	0.818182
test_2_for_all_find_SGD_pca_del	0.88889	0.829545
test_2_for_all_find_log_pca_del	0.88462	0.829545
test_2_for_all_find_forest_del_back	0.88000	0.875000
test_2_for_all_find_SVC_classic	0.88000	0.886364
test_2_for_all_find_lin_del_back	0.86957	0.840909
test_2_for_all_find_metric_classic	0.85000	0.920455

Модель метрической классификации без
обработки корреляции предикторов:
полнота 85%, точность 92%

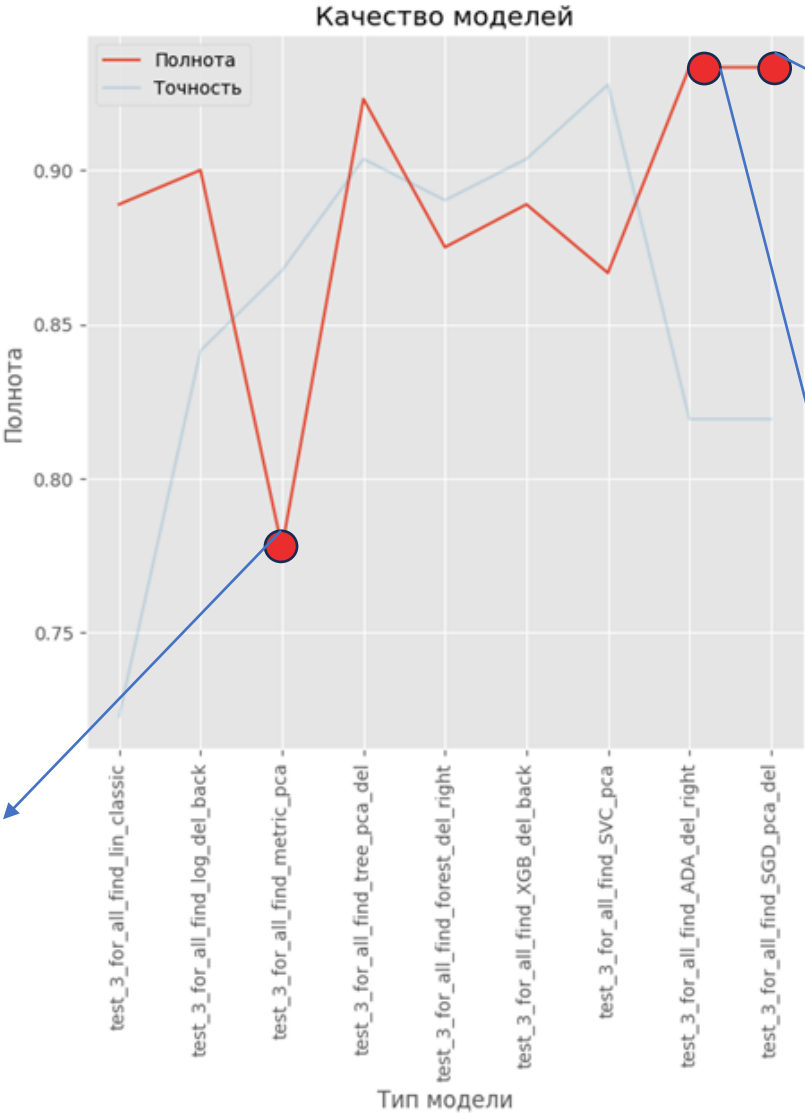


$$recall = \frac{TP}{TP + FN},$$

РЕЗУЛЬТАТЫ. ИСТОРИИ ГЛУБИНЫ 3

	pod	accur
test_3_for_all_find_ADA_del_right	0.93333	0.819277
test_3_for_all_find_SGD_pca_del	0.93333	0.819277
test_3_for_all_find_tree_pca_del	0.92308	0.903614
test_3_for_all_find_log_del_back	0.90000	0.841463
test_3_for_all_find_lin_classic	0.88889	0.722892
test_3_for_all_find_XGB_del_back	0.88889	0.903614
test_3_for_all_find_forest_del_right	0.87500	0.890244
test_3_for_all_find_SVC_pca	0.86667	0.927711
test_3_for_all_find_metric_pca	0.77778	0.867470

Модель метрической классификации с применением метода главных компонент: полнота 78%, точность 87%



Модель стохастического градиентного спуска (SGD) с удалением коррелирующих предикторов и применением метода главных компонент: полнота 94%, точность 82%

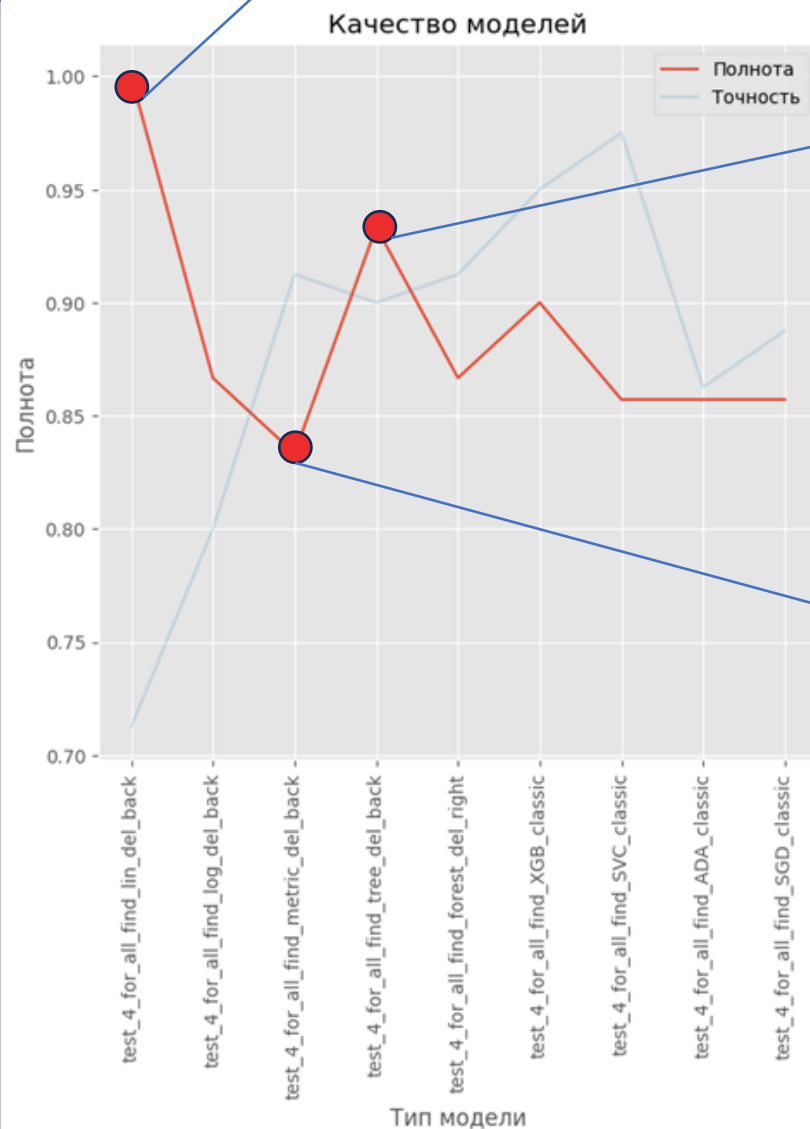
Модель адаптивного бустинга (ADA) с удалением коррелирующих предикторов: полнота 94%, точность 82%

$$recall = \frac{TP}{TP + FN},$$

РЕЗУЛЬТАТЫ. ИСТОРИИ ГЛУБИНЫ

	pod	accur
test_4_for_all_find_lin_del_back	1.00000	0.7125
test_4_for_all_find_tree_del_back	0.93333	0.9000
test_4_for_all_find_XGB_classic	0.90000	0.9500
test_4_for_all_find_log_del_back	0.86667	0.8000
test_4_for_all_find_forest_del_right	0.86667	0.9125
test_4_for_all_find_SVC_classic	0.85714	0.9750
test_4_for_all_find_ADA_classic	0.85714	0.8625
test_4_for_all_find_SGD_classic	0.85714	0.8875
test_4_for_all_find_metric_del_back	0.83333	0.9125

Модель линейной классификации
с удалением коррелирующих предикторов:
полнота 100%, точность 71%



Модель решающего дерева
с удалением коррелирующих
предикторов: полнота 93%,
точность 90%

Модель метрической
классификации с удалением
коррелирующих предикторов :
полнота 83%, точность 91%

Тестирование свойств моделей в «квазиоперативном» режиме

ОБЪЕКТ
ДАННЫЕ 2021 ГОД

	num	dt	fnum	fdt	bdt	surface	gaxe_km	paxe_km	tth	tmin
id_all										
1	1	202109021115	1	202109021100	202109021015	39446.988281	419.837982	176.138718	4.0	-7.783545
2	2	202109021115	2	202109021100	202109020915	17717.996094	300.650055	154.332443	4.0	-4.902808
3	3	202109021115	45	202109021100	202109020615	13918.201172	252.403183	93.541786	-36.0	-47.782158
4	4	202109021115	46	202109021100	202109020945	1347.320435	54.508865	38.402035	-36.0	-46.570610
5	5	202109021115	-1	-1	202109021115	1118.772827	83.267319	22.475645	-32.0	-41.135063
...	...	...	...	...	...	...	...	...	...	...
161727	115	202108110430	-1	-1	202108110430	1244.971802	61.878784	30.109180	-1.0	-8.302069
161728	116	202108110430	114	202108110415	202108110145	1158.605347	113.459312	20.482304	-1.0	-12.569494
161729	117	202108110430	110	202108110415	202108110330	426.596771	55.303505	11.542769	2.0	-5.236823
161730	118	202108110430	123	202108110415	202108110400	502.490601	49.746040	14.408597	3.0	-6.415961
161731	119	202108110430	23	202108110415	202108110345	369.483032	29.357155	14.121096	4.0	-10.408484

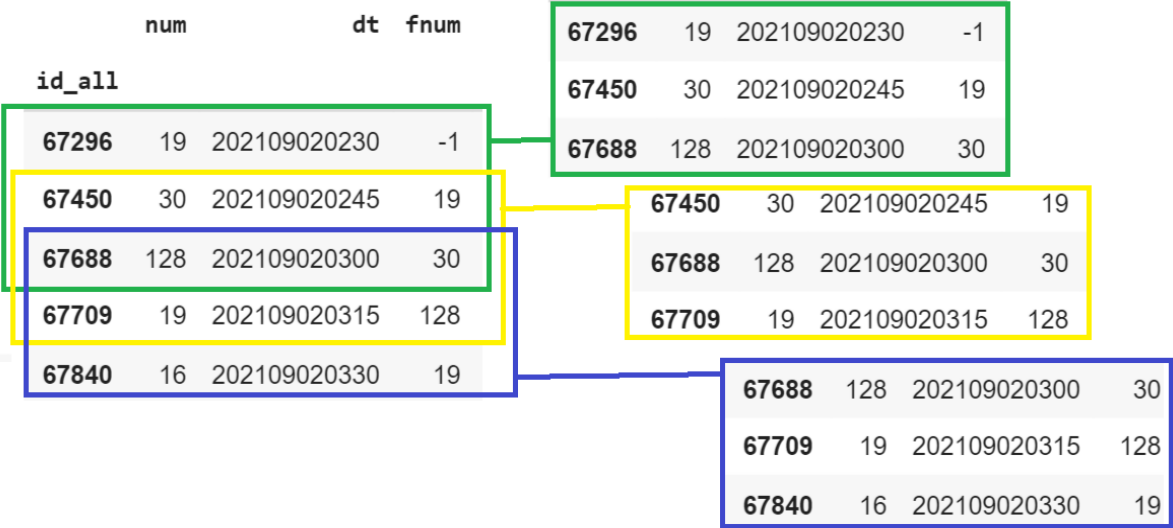
161731 rows × 60 columns

Часть DataFrame со всеми параметрами облачных ячеек за 2021 год

	num	dt	bnum	bdt
id				
1	75	202107040830	57	202107040645
2	67	202107220545	40	202107220330
3	17	202107231130	49	202107230930
4	59	202108080715	54	202108080300
5	46	202108081015	106	202108080745
6	103	202108101415	101	202108101330
7	74	202108110830	80	202108110545
8	41	202108130630	115	202108121730
9	109	202108131315	115	202108121730
10	21	202108131400	115	202108121730
11	44	202108191045	33	202108182230
12	28	202108310645	27	202108310615
13	13	202109010815	9	202109010730
14	91	202109020545	19	202109020230
15	61	202109120500	66	202109120300
16	3	202109250930	86	202109250900
17	89	202110181100	107	202110180630

Зафиксированные смерчевые измерения за 2021г

Определение прогностных свойств избранных моделей на новых данных



Пример построения подвыборок глубиной 3 из полной истории смерчевой ячейки (2021 год)

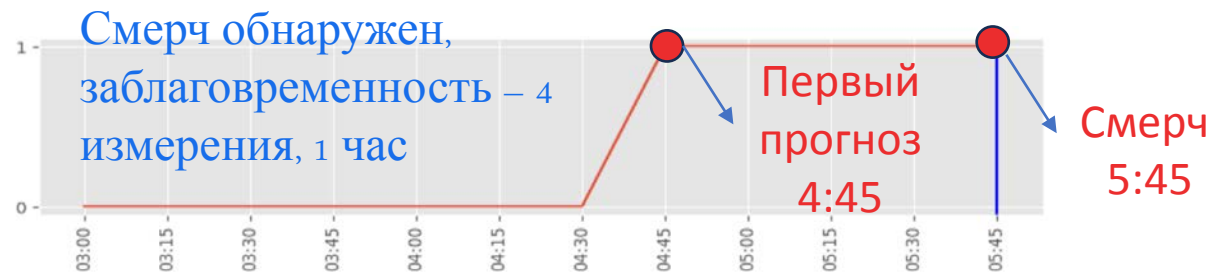
Тестовые данные	1/0	1/0	1/0	1/0	1/0	1/0	1/0	смерч
Классификация	1	0	0	0	1	0	1	1
Заблаговременность - 7 измерений								

- **Заблаговременность** – отнесение облачной ячейки к классу смерчевых до возникновения смерча (возможность прогноза)
- **Факт обнаружения** – умение моделей распознавать ячейки со смерчами (классификация)
- **Доля ложно-смерчевых ячеек** – отнесение безопасной облачной ячейки к классу смерчевых

```
test_3_for_all_find_ADA_del_right
Тестовые ответы [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Предсказания [0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Успех
```

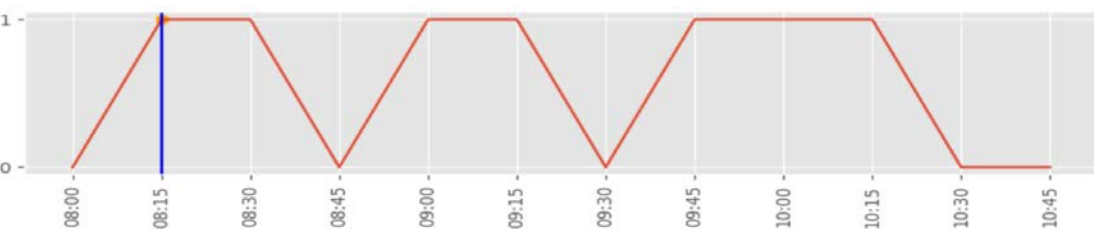
```
test_3_for_all_find_ADA_del_right
Тестовые ответы [0. 0. 0. 0. 0. 0. 0.]
Предсказания [0. 0. 0. 0. 0. 1. 0.]
Ложное распознавание или вероятность смерча
Успех
```

```
test_3_for_all_find_ADA_del_right
Тестовые ответы [0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
Предсказания [0. 0. 0. 0. 0. 1. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
Ложное распознавание или вероятность смерча
Неудача
```



Смерч: 02.09.2021, 05:45

Смерч обнаружен

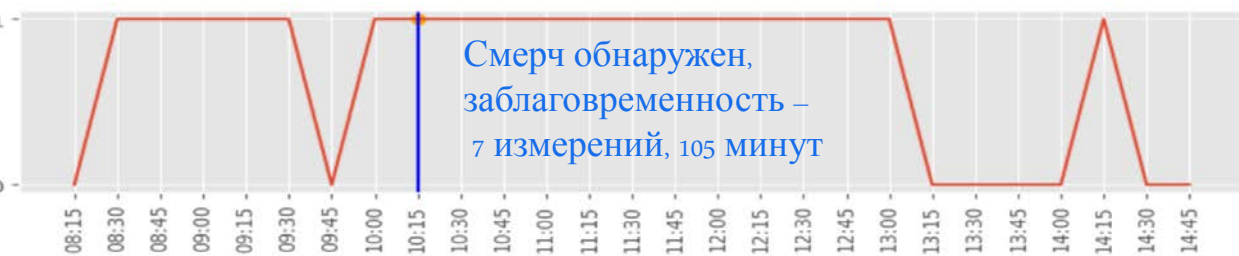


Смерч: 01.09.2021, 08:15

Предсказан за 8 измерений, 2 часа



Смерч: 11.08.2021, 08:30



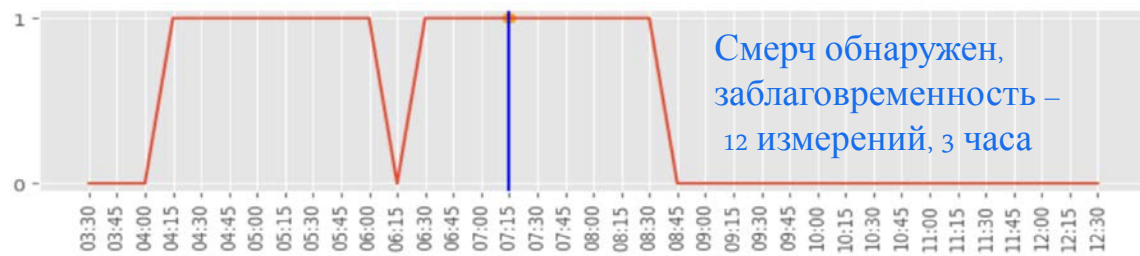
Смерч: 08.08.2021, 10:15

Динамики спрогнозированных классов опасности смерчей

МОДЕЛЬЮ

`test_3_for_all_find_ADA_del_right`

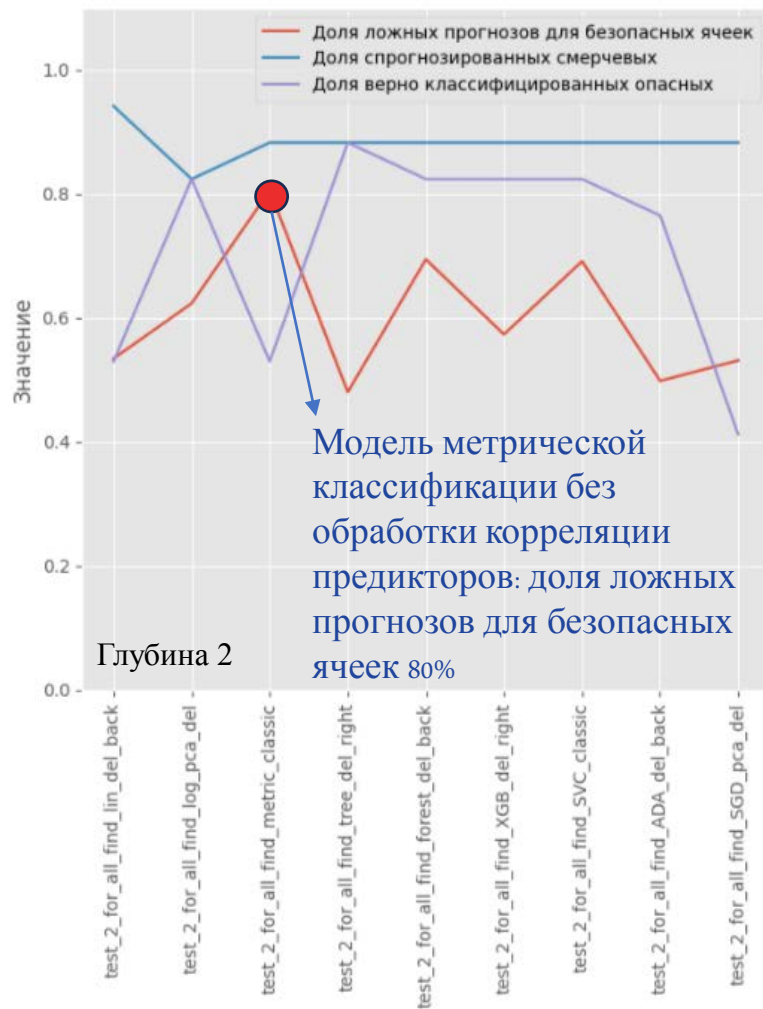
Таким образом в ходе комплексного анализа были получены следующие оценки заблаговременной предупрежденности смерчей моделями *history_type = 3*): *lin* - 94%, *log* - 77%, *SVC* - 82% *SGD* - 82%, *metric* - 70%, *tree* - 82%, *forest* - 76%, *XGB* - 71%, *ADA* - 76%.



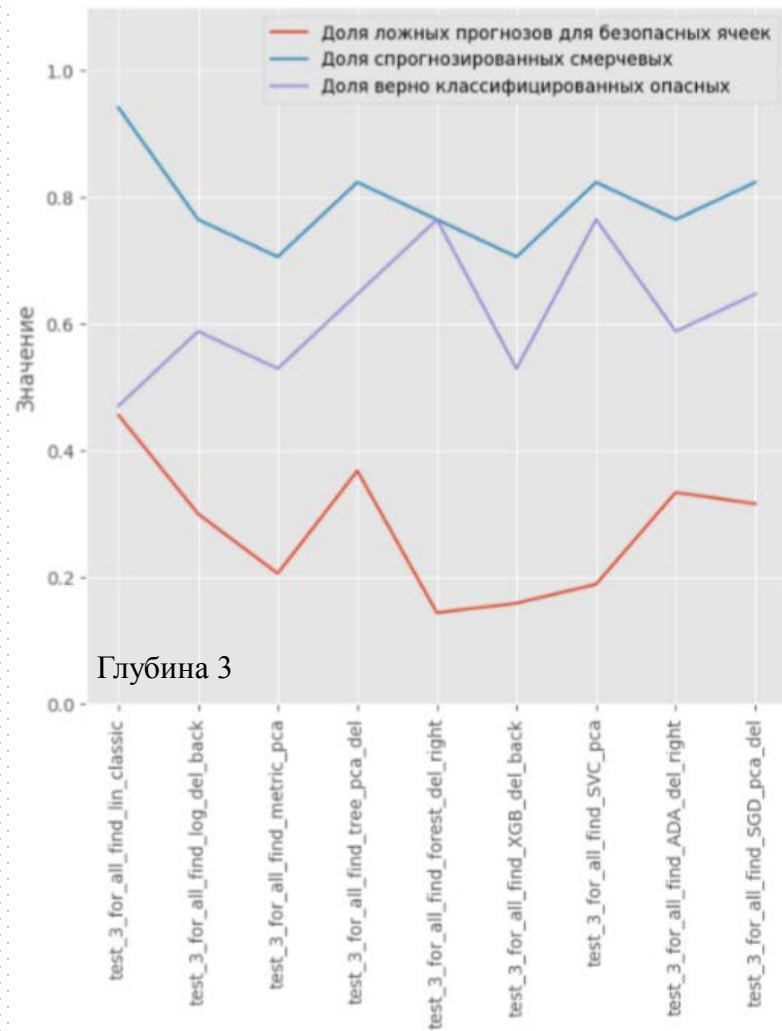
Смерч: 08.08.2021, 07:15

Качество прогнозных свойств избранных моделей

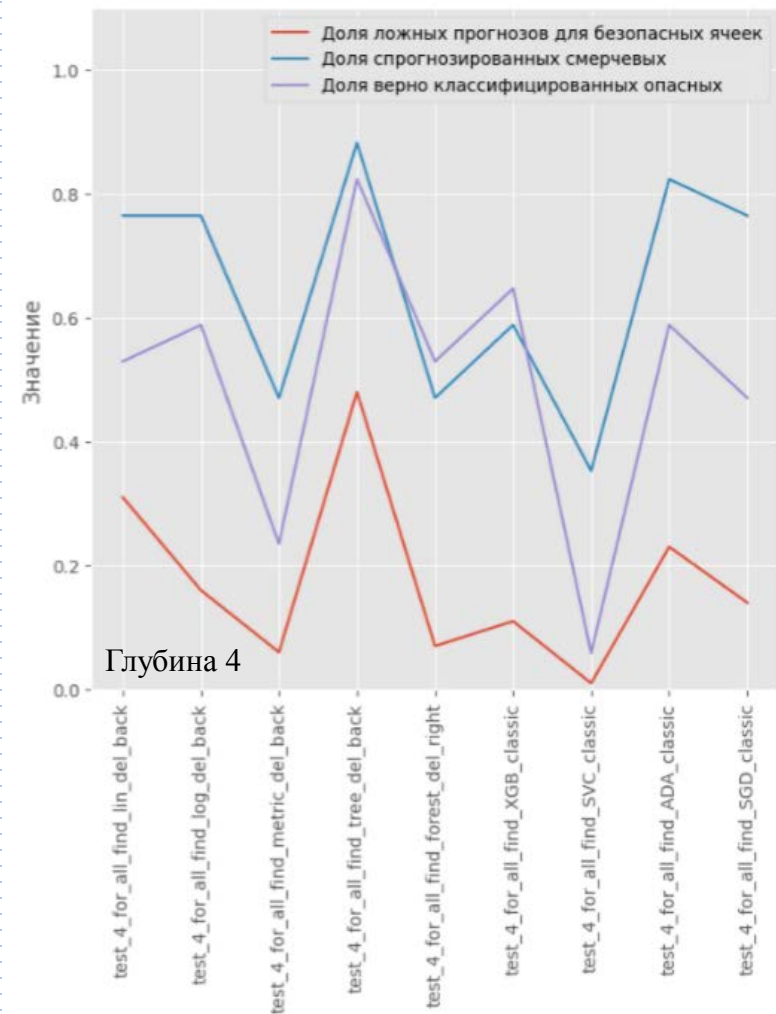
Качество прогнозных свойств



Качество прогнозных свойств



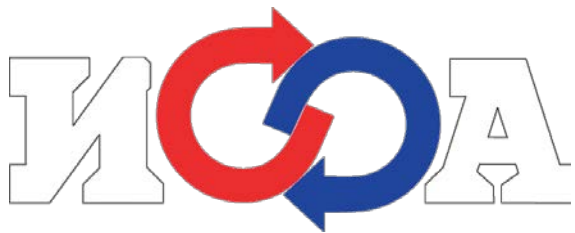
Качество прогнозных свойств



РЕЗУЛЬТАТЫ

РАБОЧИЕ МОДЕЛИ КЛАССИФИКАЦИИ

Получено семейство моделей классификации облачных ячеек для распознавания смерчеобразующих облаков над Черным морем. В ходе анализа отобраны наиболее подходящие модели с лучшими показателями по метрикам качества.



МЕТОДИКА ПОДГОТОВКИ ДАННЫХ

Разработана методика обработки динамических характеристик облачных ячеек, построены и отсортированы по степени важности предикторы моделей

ПРОГНОСТИЧЕСКИЕ СВОЙСТВА МОДЕЛЕЙ

Оценены прогнозные свойства полученных моделей классификации на новых данных. Заблаговременность прогнозирования для большинства моделей превышает 80%.

РАЗНООБРАЗИЕ И ВАРИАТИВНОСТЬ

Построены модели строились для историй различной глубины. Такой подход позволяет расширить область покрытия прогноза в реальных условиях и производить перестройки моделей по мере поступления новой информации.

Для повышения точности прогноза проводится исследование по комбинированному применению различных моделей для принятия решения.

ПЕРСПЕКТИВЫ

В перспективе планируется усовершенствование моделей путем накопления новых данных и интеграции радиолокационных данных. Планируется внедрение глубокого обучения с использованием нейронных сетей для извлечения новых зависимостей из расширенного объема данных.

Планируется уделить особое внимание физическим аспектам процесса для верной интерпретации результатов исследования и для выявления по ним скрытых особенностей конвективного явления.

Спасибо за внимание!